Missense Mutations and Evolutionary Conservation of Amino Acids: Evidence That Many of the Amino Acids in Factor IX Function as "Spacer" Elements[1]

Cynthia D. K. Bottema,* Rhett P. Ketterling,* Setsuko Ii,* Hong-Sup Yoon,*
John A. Phillips III,† and Steve S. Sommer*

*Department of Biochemistry and Molecular Biology, Mayo Clinic/Foundation, Rochester, MN; and †Division of Genetics, Department of Pediatrics, Vanderbilt University School of Medicine, Nashville

Summary

We report 31 point mutations in the factor IX gene and explore the relationship between the level of evolutionary conservation of an amino acid and the probability of a mutation causing hemophilia B. From our total sample of 125 hemophiliacs and from those reported by others, we identify 95 independent missense mutations, 94 of which occur at amino acids that are evolutionarily conserved in the available mammalian factor IX sequences. The likelihood of a missense mutation causing hemophilia B depends on whether the residue is also conserved in the factor IX–related proteases: factor VII, factor X, and protein C. Most of the possible missense mutations in generically conserved residues (i.e., those conserved in factor IX and in all the related proteases) should cause disease. In contrast, missense mutations in factor IX–specific residues (i.e., those conserved in human, cow, dog, and mouse factor IX but *not* in the related proteases) are sixfold less likely to cause disease. Missense mutations at nonconserved residues are 33-fold less likely to cause disease. At least three models are compatible with these observations. A comparison of sequence alignments from four and nine species of factor IX and an examination of the missense mutations occurring at CpG residues suggest a model in which most residues fall on opposite ends of a spectrum. In about 40% of residues, virtually any missense mutation in a minority of the residues will cause disease, while virtually *no* missense mutations will cause disease in most of the remaining residues. Thus, many of the residues in factor IX are spacers; that is, the main chains are presumably necessary to keep other amino acid interactions in register, but the nature of the side chain is unimportant.

Introduction

Factor IX is a coagulation serine protease zymogen with eight functional domains, including (1) a signal peptide, (2) a pro-peptide which is necessary for the γ-carboxylation of the mature protein, (3) a gla domain with 12 γ-carboxyglutamic (gla) residues which bind four to six molecules of calcium, (4) a short aromatic amino acid stack, (5) a first epidermal growth factor domain which contains a high-affinity calcium-binding site, (6) a second epidermal growth factor domain of unknown function, (7) an activation peptide which is removed during proteolysis by factors VII or XI, and (8) a catalytic domain which activates factor X (reviewed in Furie and Furie 1988). Factor IX, factor VII, factor X, and protein C are closely related coagulation serine proteases that have the same eight functional domains and similar gene structure (Furie and Furie 1988).

Since the factor IX gene is located on the X chromosome, a mutation that disrupts function affects any male who receives that allele. Many mutations in the factor IX gene causing hemophilia B have been described (reviewed in Giannelli et al. 1990). The mutation rate is dramatically enhanced at CpG dinucleotides (Koeberl et al. 1989; Green et al. 1990) but not at any other dinucleotides (Bottema et al. 1991).

Received April 18, 1991; final revision received June 17, 1991.
Address for correspondence and reprints: Dr. Steve S. Sommer, Department of Biochemistry and Molecular Biology, Mayo Clinic/Foundation, Rochester, MN 55905.

[1] This paper is dedicated to Harold and Hadassah Sommer.

0002-9297/91/4904-0015$02.00

Herein we present 31 new families with point mutations and analyze the relationship between evolutionary conservation of amino acids and missense mutations which cause hemophilia B in humans.

Methods

Sequencing

DNA was extracted from blood collected in ACD solution B or solution A as previously described (Gustafson et al. 1987). Regions of likely functional significance were sequenced by *g*enomic *a*mplification *w*ith *t*ranscript *s*equencing (GAWTS) (Stoflet et al. 1988) as described by Sommer et al. (1990). GAWTS is a method of direct sequencing that involves (1) PCR amplification of the segment of interest, where at least one of the PCR primers has an attached phage promoter sequence, (2) transcription of the amplified segment with the phage RNA polymerase to produce a single-stranded RNA molecule, and (3) sequencing of the RNA template with reverse transcriptase. The following bases were sequenced (numbering system corresponds to that of Yoshitake et al. [1985]): region A, −106 to 139; region B/C, 6720 to 6265; region D, 10544 to 10315; region E, 17847 to 17601; region F, 20577 to 20334; region G, 30183 to 29978; and region H, 31411 to 30764. The poly A addition region was not sequenced. The order of the numbers in each region indicates the direction of sequencing. In all, at least 2.2 kb was sequenced from each hemophiliac.

Haplotype Analysis

The following polymorphisms in the factor IX gene were examined: *Hin*fI (intron a) (Winship et al. 1984), *Xmn*I (intron c) (Winship et al. 1984), *Taq*I (intron d) (Camerino et al. 1984), and *Hha*I (3′ of gene) (Winship et al. 1989). From these four polymorphisms, eight common haplotypes were defined, with frequencies of 2%–19% in the normal Caucasian population (Ketterling et al., in press).

DNA segments containing the *Taq*I and the *Xmn*I restriction sites were amplified by PCR and digested with the appropriate restriction enzyme as described elsewhere (Koeberl et al. 1990). The products were gel electrophoresed, and the presence (+) or the absence (−) of the restriction site was determined. For the *Hin*fI (also known as *Dde*I) polymorphism, the DNA was amplified by PCR, and the presence (+) or absence (−) of the 50-bp insert was determined by gel electrophoresis. The *Hha*I polymorphism was determined by amplifying 500 ng genomic DNA by PCR using 0.1 μM of the previously described oligonucleotides H1 and H2 (Winship et al. 1989) and 1.5 mM $MgCl_2$ in 50-μl reactions. The PCR products were digested with *Hha*I, and the presence (+) or absence (−) of the restriction site was determined by gel electrophoresis.

Levels of Amino Acid Conservation in Factor IX

Four classes of residues can be defined from the available factor IX sequences and from the sequences of both human and bovine factor VII, factor X, and protein C (fig. 1 and Appendixes A and B). A residue is "generic" if it is identical in all species of factor IX and is also identical in the three related blood coagulation serine proteases: factor VII, factor X, and protein C. The residue is "factor IX specific" if it is identical in all species of factor IX but *not* identical in any of the three related proteases. The residue is "partially generic" if it is identical in all species of factor IX and is identical in one or two of the three related proteases. If a residue is conservatively substituted in the species of factor IX (i.e., is S/T, S/A, Y/F, R/K, I/L/V, N/D, D/E, Q/N, and E/Q), the above definitions are modified to also allow the conservative substitution in the three related proteases. If a residue is nonconservatively substituted in any of the species of factor IX, it is classified as "nonconserved."

The above definitions differ from a previous classification (Koeberl et al. 1990; Sarkar et al. 1990) in that the sequence alignment utilizes the complete sequence of dog factor IX (Evans et al., 1989) and mouse factor IX (Wu et al. 1990) and bovine factor VII (Takeya et al. 1988) (Appendixes A and B). Most important, the presently defined factor IX–specific residues and most of the partially generic residues were previously combined into one class.

A detailed protocol was used for assigning residues to each class (see Identity subsection). Our conclusions will remain the same despite certain revisions in the classification protocol (see Conservative substitutions subsection).

1. *Identity.* — Those amino acids identical in the mammalian factor IX sequences were compared with the homologous residues in the related coagulation serine proteases. The amino acid was assigned to a class on the basis of extent of identity with these proteases. For a residue to be considered conserved in a given related protease, both the human and bovine residues needed to be identical with the corresponding factor IX amino acid.

2. *Conservative substitutions.* — Each amino acid which was not identical in the factor IX sequences was considered to be nonconserved unless the substitutions were highly conservative. These highly conservative substitutions were S/T, S/A, E/D, D/N, Q/E, Q/N, F/Y, K/R, and I/L/V (see fig. 1 legend for the single-letter amino acid code). Conservative substitutions were defined rather stringently in that, with one exception (I/L/V), only two residues constitute each conservation group. As examples, D (aspartate) and N (asparagine) are in one group because they are related polar residues of approximately the same volume. D and E (glutamate) are in another group because they both have a negative charge. Thus, at any given D, the conservative substitutions were limited to *either* charge or size. Therefore, the presence of D and E in the factor IX sequences is classified as a conservative substitution. While the presence of D, E, and N is classified as nonconservative. (The alternative possibility of allowing *any* combination of either (*a*) D, E, Q, and N or (*b*) S, T, and A converts only four nonconserved residues to the conserved class and does not alter any of the conclusions). The conservatively substituted factor IX amino acids were then compared with the related serine proteases and were assigned to one of the above classes (i.e., generic, partially generic, or factor IX specific).

In practice, almost all of the generic, partially generic, and factor IX–specific residues are identical rather than conservatively substituted. Conservative substitutions in factor IX occur in only 8% of the 364 conserved amino acids. If conservative substitutions are not allowed, the number of residues involved in hemophilia B remains virtually unchanged (182 amino acids without substitutions vs. 176 amino acids with conservative substitutions [table 3]). However, the number of mutations in nonconserved residues increases from one to seven if conservative substitutions are not allowed. Mutations were observed at one N/D site, four I/L/V sites, one T/S conservative site (see fig. 1).

3. *Mutations at CpG versus non-CpG sites.* — Since mutations at the dinucleotide CpG occur more frequently than those at other sites (Koeberl et al. 1989; Green et al. 1990), the mutations at CpG and non-CpG nucleotides were analyzed separately (see tables 3 and 5). Residues at CpG dinucleotides were assigned to the non-CpG or CpG categories. This assignment was based on the fraction of possible mutations at CpG and non-CpG nucleotides in each residue (table 2). As examples, all of the arginine residues with codons of CGX were assigned to the CpG dinucleotide group. However, a glycine residue preceded by a residue ending in C (XX$\underline{C}$ $\underline{G}$GX) had the first G assigned to the CpG group and the second G assigned to the non-CpG group. Although one-third of all independent mutations occur at CpG, the rarity of this dinucleotide in the factor IX coding sequence stipulates that the CpG group accounts for less than 3% of all the possible kinds of missense mutations.

Results

Mutations

Point mutations were delineated in 31 families with hemophilia B by direct genomic sequencing of the regions of likely functional significance, which include the coding region, the splice junctions, the putative promoter, the 5′ untranslated region, and a small part of the 3′ untranslated region (Koeberl et al. 1989). In total, 66 kb of sequence were obtained. While the majority of the hemophiliacs are Caucasians of northern-European descent, two (HB101 and HB102) are Hispanic and one (HB109) is Japanese.

Of the 31 point mutations, two affect splice junctions and six produce nonsense mutations, but the great majority (23) are missense mutations (table 1). Only one sequence change was found in each individual. The splice junction mutations which disrupt known consensus sequence, as well as the nonsense mutations which result in truncated protein products, are clearly causative mutations. In addition, we conclude that the missense mutations are also all causative because (1) they are the only sequence change found in the regions of likely functional significance, (2) polymorphisms in these regions are rare (Koeberl et al. 1989), (3) these changes are not present in normal individuals or as second site changes in other hemophiliacs, and (4) the missense mutations (except for one) are all at evolutionarily conserved residues (fig. 1).

Twenty-two of these mutations have not been previously described. Two mutations (serine^{94} and lysine^{412}) occur at residues specifically conserved in factor IX but not in the related proteases (fig. 1). Thus, these mutated residues may be important for factor IX–specific interactions such as binding to factors VIII or VII. Asparagine^{347}→isoleucine (HB108) is the first reported missense mutation in a hemophiliac to occur in a nonconserved residue. Tryptophan^{407}→arginine represents the first non-CpG site at which two patients (HB20 and HB92) have the same mutation in a differ-

Table I

Single Base Mutations in Factor IX

Family	Factor IX Coagulant Reported (%)	Nucleotide Change	Nucleotide Number	Structural Change	Domain	CpG Mutation[a]	Conservation Class[b]	Previous Report
HB64	<1	G→A	117	V^{-17}→I	pro	No	P	
HB90	4[c]	C→T	6364	R^{-4}→W	pro	Yes	P	Giannelli et al. 1990
HB116	<1	G→A	6428	C^{18}→Y	gla	No	G	
HB97	<1	G→T	10406	E^{52}→TAG	EGF 1	No		
HB106	1	A→T	17697	R^{94}→S	EGF 2	No	S	
HB111	<1[d]	T→C	17743	S^{110}→P	EGF 2	No	P	
HB88	2	G→A	17786	C^{124}→Y	EGF 2	No	G	
HB68	1	G→A	17797	V^{128}→M	EGF 2	No	P	
HB115	<1	G→A	20375	C^{132}→Y	EGF 2	No	G	
HB120	11	G→A	20414	R^{145}→H	Activation peptide	Yes	P	Giannelli et al. 1990, Koeberl et al. 1990
HB102	<1	G→C	30038	Intron f, − 1	Splice acceptor	No		
HB65	15	T→G	30101	I^{216}→M	Catalytic	No	G	
HB114	<1	G→A	30821	Intron g, − 1	Splice acceptor	No		Giannelli et al. 1990
HB98 and HB104 ...	6 and 15	G→A	30864	R^{248}→Q	Catalytic	Yes	P	Giannelli et al. 1990, Koeberl et al. 1990
HB91	<1[d]	C→T	30875	R^{252}→TGA	Catalytic	Yes		Giannelli et al. 1988
HB96	<1	T→C	30930	I^{270}→T	Catalytic	No	P	
HB101	8[c]	T→C	30945	L^{275}→P	Catalytic	No	G	Giannelli et al. 1990
HB109	<1[d]	T→G	30985	I^{228}→M	Catalytic	No	P	
HB100	<1	G→T	31001	E^{294}→TAA	Catalytic	No		
HB82	4	C→T	31091	Q^{324}→TAG	Catalytic	No		
HB122	4	C→G	31096	Y^{325}→TAG	Catalytic	No		
HB105..................	4	C→T	31118	R^{333}→TGA	Catalytic	Yes		Giannelli et al. 1990; Koeberl et al. 1990
HB110	<1	G→T	31118	R^{333}→L	Catalytic	Yes	P	
HB107	2	G→A	31119	R^{333}→Q	Catalytic	Yes	P	
HB108	5[d]	A→T	31161	N^{347}→I	Catalytic	No	N	
HB124	<1	G→A	31165	M^{348}→I	Catalytic	No	G	
HB87	<1	G→A	31218	G^{366}→E	Catalytic	No	G	
HB86	<1[d]	A→G	31281	E^{387}→G	Catalytic	No	P	
HB92	<1	T→C	31340	W^{407}→R	Catalytic	No	G	Koeberl et al. 1989
HB66	3	C→A	31356	T^{412}→K	Catalytic	No	S	

[a] Mutations at the dinucleotide CpG. All but R^{333}→L were transitions.

[b] See Methods for definitions. G = conserved; P = partially generic; S = factor IX specific; and N = nonconserved.

[c] Posttherapy values but clinically severe by criteria of Eyster et al. (1980).

[d] Values measured by our laboratory. Otherwise, values are as reported by the referring hemophilia centers.

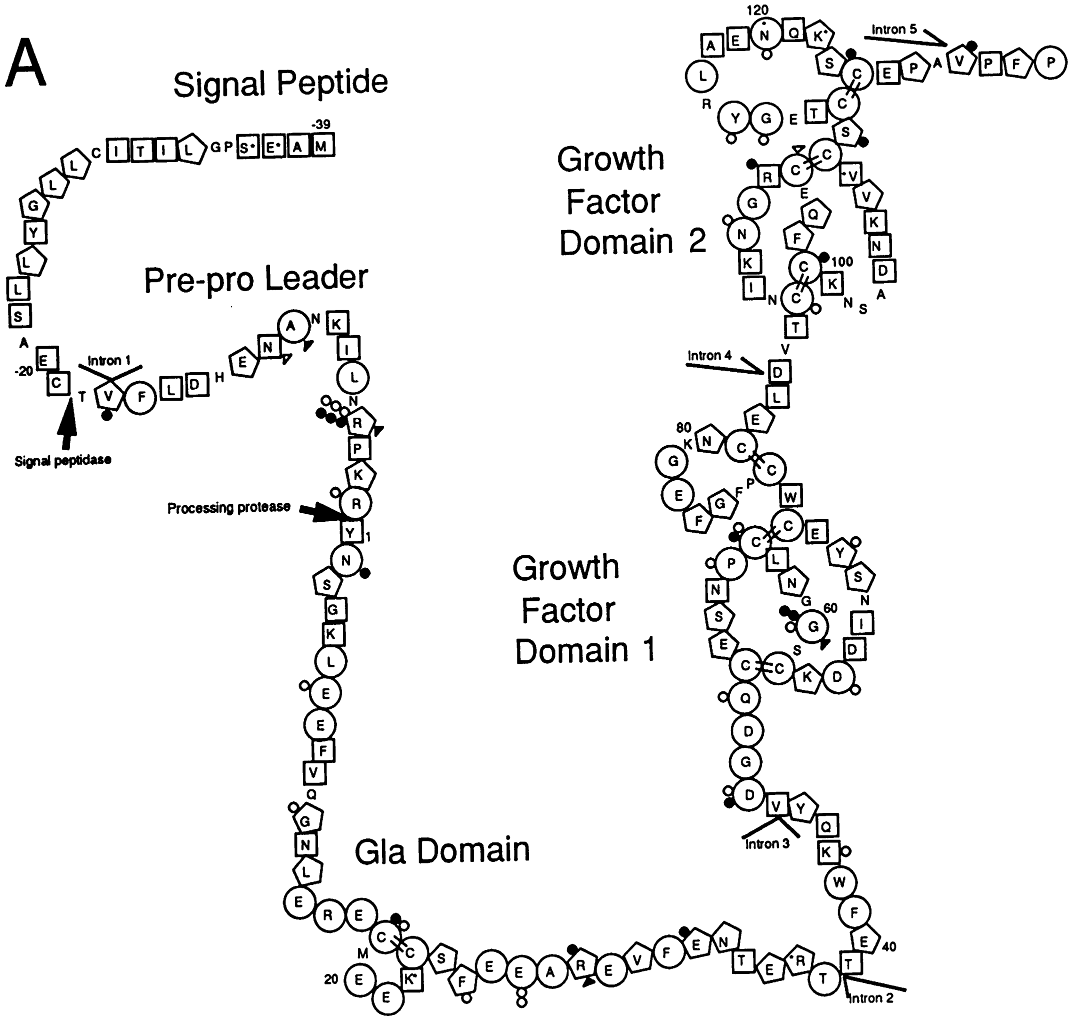
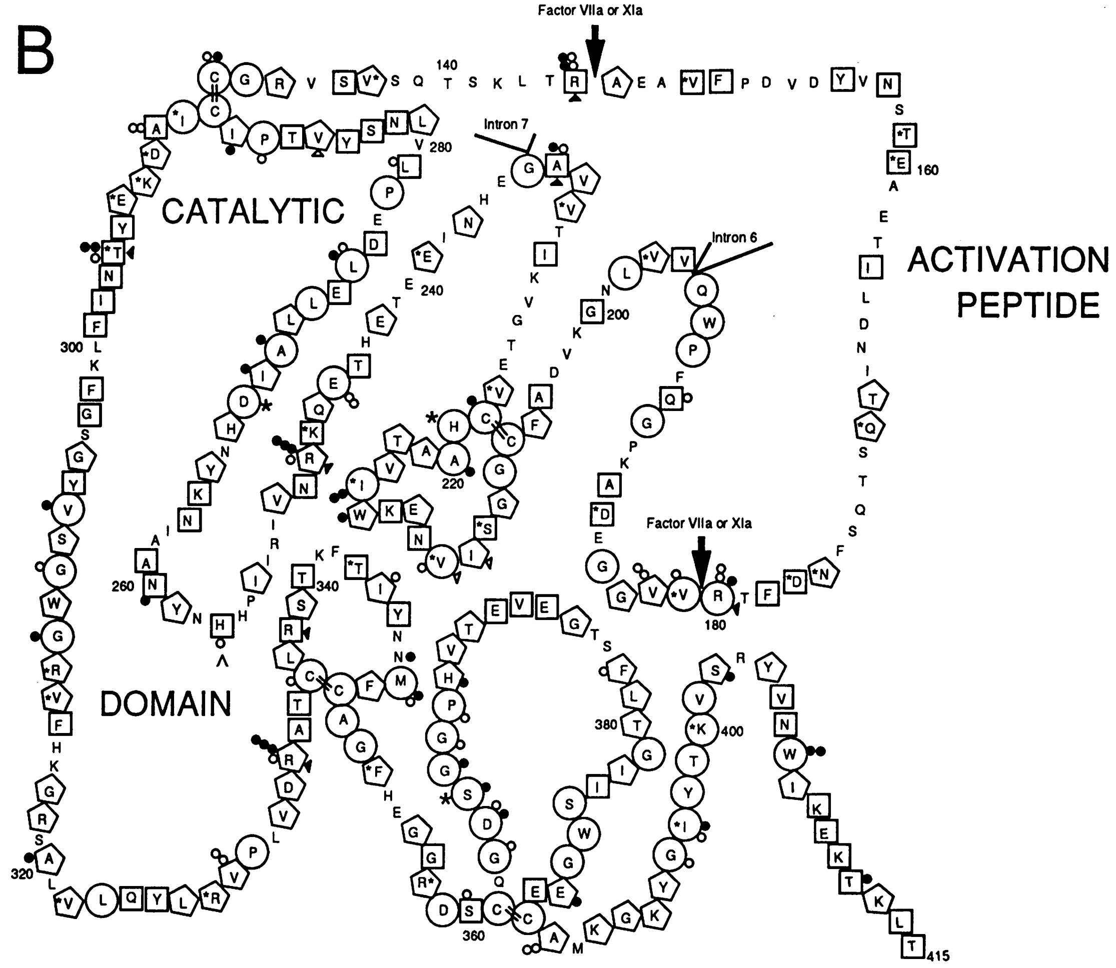


Figure 1 Factor IX missense mutations and amino acid conservation. Both the conservation of the amino acid residues in factor IX and the location of missense mutations are shown. ● = Missense mutation at a given residue from our sample; ○ = missense mutation at a given residue reported by others (Giannelli et al. 1990). Multiple symbols (● or ○) indicate the number of known independent missense-mutation changes at a given residue. For CpG dinucleotides, symbols are as follows: ▲ = conserved residues in which transitions *and* transversions will cause a nonconservative missense substitution; △ = conserved residues where *only* transversions will cause a nonconservative missense substitution. Λ = Missense mutation not causing disease (Montandon et al. 1990). The geometric shapes indicate the degree of conservation of an amino acid residue, as determined from factor IX and related serine protease alignments (Appendixes A and B). Circles represent "generic" residues, squares represent "factor IX-specific" residues, and pentagons represent "partially generic" residues as defined in Methods. An asterisk (*) inside the geometric shape indicates that conservative substitutions occurred in the mammalian species of factor IX. The alignment of factor IX is based on the available mammalian species of factor IX and on both the human and bovine sequences of factor X, factor VII (Takeya et al. 1988), and protein C (Koeberl et al. 1990; Sarkar et al. 1990) (see Appendixes A and B). *A*, Alignment of N-terminal segment of factor IX, based on four available mammalian species: human (Yoshitake et al. 1985), cow (Katayama et al. 1979), dog (Evans et al. 1989), and mouse (Wu et al. 1990). The translation start site is based on additional data from rat and macaque sequences (Pang et al. 1990). *B*, Alignment of C-terminal segment of factor IX (activation peptide and catalytic domain), based on nine mammalian species: human, sheep, pig, rabbit, guinea pig, rat, mouse, cow, and dog (Evans et al. 1989; Sarkar et al. 1990). The single-letter code for amino acids is as follows: A = Ala; R = Arg; N = Asn; D = Asp; C = Cys; Q = Gln; E = Glu; G = Gly; H = His; I = Ile; L = Leu; K = Lys; M = Met; F = Phe; P = Pro; S = Ser; T = Thr; W = Trp; Y = Tyr; and V = Val. Stars (★) indicate the serine protease catalytic triad amino acids. Note that the classification in table 3 is based on factor IX from only four species and that only non-CpG residues were tabulated.



ent haplotype, indicating that these two mutations are due to independent events.

Missense Mutations versus Extent of Evolutionary Conservation

To determine the relationship between missense mutations in hemophiliacs and the amino acid conservation classes, haplotype analysis was used to determine whether recurrent mutations were independent events. Mutations at non-CpG dinucleotides were considered separately from those at CpG dinucleotides, since mutations at CpG dinucleotides are greatly enhanced. Although the CpG category represents less than 3% of all the missense mutations that are possible, the analysis of this category is important (see below).

At least 38 independent missense mutations were found at non-CpG residues (table 1; Bottema et al. 1990*a;* Koeberl et al. 1990). We combined these data with data on 43 additional non-CpG mutations reported in the literature (reviewed in Giannelli et al. 1990). Since haplotype data are not generally available on patients reported by others, it was not possible to identify recurrent mutations that are due to independent mutational events. However, in our sample, extensive haplotype analysis strongly suggests that only one of the 25 recurrent mutations at *non-CpG* dinucleotides is independent (Bottema et al. 1989, 1990*b;* Ketterling et al. 1991, and in press). Thus, it seems reasonable to assume that all recurrent mutations reported in the literature are the result of common ancestors. Common ancestry has been well documented for the most frequently recurring non-CpG mutation: isoleucine[397]→threonine (Bottema et al. 1990*b;* Thompson et al. 1990).

Four classes of factor IX residues can be defined on the basis of the extent of evolutionary conservation. These classes are (1) generic residues that are conserved in factor IX and in the related coagulation serine proteases factor VII, factor X, and protein C, (2) partially generic residues that are conserved in one or two of the related coagulation proteases, (3) factor IX–specific residues that are conserved in factor IX but in none of the related coagulation proteases, and (4) nonconserved residues (see Methods).

The amino acids in factor IX are distributed between the four classes in a nonrandom manner (table 2). Almost one-half of the generic residues are either cysteine, glycine, or glutamate. The cysteine residues are involved in disulfide bonds (fig. 1), and most of these glutamate residues are modified to γ-carboxyglutamic acids necessary for calcium binding (Furie and Furie 1988) (table 2). A substantial fraction of the charged generic residues (10 glutamates and three aspartates) are involved in the chelation of calcium (Rees et al. 1988). If calcium binding is ignored, charged residues are substantially underrepresented in the generic and partially generic classes. In contrast, codons having a high G + C content are substantially overrepresented (Bottema et al. 1991). If the generic and partially generic residues are combined, cysteine, glycine, and tryptophan are significantly overrepresented while threonine, asparagine, and lysine are underrepresented (for χ^2 values, see table 2, footnotes b and c).

Both the classification of each residue in factor IX and the location of all the missense mutations observed in patients with hemophilia B were determined (fig. 1). The non-CpG missense mutations are distributed throughout the factor IX protein. However, missense mutations causing hemophilia B are most likely to occur at generic residues (table 3). Mutations at the partially generic residues are about twofold less likely to produce hemophilia B; and mutations at the factor IX–specific residues are sixfold less likely to produce hemophilia B (i.e., the likelihood that they will produce the disease is only 15% of that for generic residues) (table 3). Mutations at nonconserved residues are 33-fold less likely to produce hemophilia B, indicating that these residues are almost never involved in this disease.

Table 2

Number of Residues in Each Class for Human Factor IX

Residue Type[a]	No. of Generics	No. of Partial Generics	No. of Factor IX Specifics	No. of Nonconserved
Nonpolar:				
A	5	3	9	7
V	4	17	7	8
L	7	9	7	5
I	3	7	9	5
P	7	1	2	5
F	3	8	6	4
W[b]	4	2	1	0
M	1	0	1	2
Subtotal	34	47	42	36
Polar:				
G[b]	19	10	4	3
S	2	10	5	9
T[c]	2	4	14	10
C[b]	22	0	1	1
Y	2	6	8	0
N[c]	3	5	12	11
Q	2	3	4	4
Subtotal	52	38	48	38
Charged:				
D	6	2	7	4
E(γ)[d]	11 (9)	10 (3)	11 (0)	11 (0)
K[c]	1	7	12	8
R	3	10	3	3
H	1	2	1	6
Subtotal	22	31	34	32
Total	108	116	124	106

Source.—Fig. 1.

[a] As categorized by Lehninger (1975).

[b] More abundant in generic and partially generic classes than in factor IX–specific and nonconserved classes. By the χ^2 test, $P < .001$ for cysteine and glycine and $P < .05$ for tryptophan.

[c] More abundant in factor IX–specific and nonconserved classes than in generic and partially generic classes. By the χ^2 test, $P < .001$ for threonine and $P < .03$ for asparagine and lysine.

[d] γ = γ-carboxyglutamic acid. The numbers in parentheses are the numbers of modified residues.

Three models for these observations can be envisioned. These models will be stated in the context of possible explanations for the low frequency of causative missense mutations at factor IX–specific residues:

1. At 15% of the factor IX–specific residues, *most,* if not all, of the possible amino acid substitutions will cause disease. The remaining 85% of the factor IX–specific residues are not essential. *In this case, factor IX sequence from additional nonmammalian species should allow sufficient evolutionary*

Table 3

Frequency of Missense Changes in Factor IX That Are Due to Mutations at Non-CpG Dinucleotides, as Function of Amino Acid Conservation

	Distribution, by Amino Acid Conservation Class, Based on Factor IX Sequences from Four Species[a]													
	Generic			Partially Generic			Factor IX Specific				Nonconserved			
	N Terminal	C Terminal	Total	N Terminal	C Terminal	Total	N Terminal	C Terminal	Total	Total Conserved	N Terminal	C Terminal	Total	Total
A. No. of residues	52	50	102	38	83	121	54	87	141	364	25	53	78	442
B. No. of missense mutations observed[b]	23	24	47	7	16	23	2	8	10	80	0	1	1	81
C. Frequency relative to generics[c]			1.0			.41			.15	...			.03	...
D. Maximal target size[d] (A × C)			102			51			21	174 (48%)			2	176 (40%)[e]

NOTE.—The signal-through-EGF2 domains of factor IX constitute the N-terminal segment (N terminal), and the activation and catalytic domains of factor IX constitute the C-terminal segment (C terminal). Factor IX residues in four species were compared with human and bovine factor VII, human and bovine factor X, and human and bovine protein C (Appendix B). The four species of factor IX included human, mouse, cow, and dog (Appendix A).

[a] See Results for definitions of the residue classes, and see Methods for details of classification (93% of the conserved residues are identical in the four species of factor IX, and 7% are conservatively substituted).

[b] Data include our missense mutations at non-CpG nucleotides of factor IX, as well as those reported by Giannelli et al. (1990). (Five mutations [6%] occur at conservatively substituted residues.)

[c] Frequencies are normalized with respect to missense mutations/generic residues.

[d] The number of residues predicted to cause hemophilia B, if it is assumed that 100% of missense mutations at generic residues cause disease (row A × Row C).

[e] When the C-terminal classification is based on nine rather than on four mammalian factor IX sequences, the target size decreases only slightly, to 39% of total residues.

time for most nonessential residues to be substituted, while the essential 15% of functionally important residues will remain conserved.

2. Mutations at *100%* of the factor IX–specific residues can cause disease, but, on average, at any given site, only 15% of the 19 possible amino acid substitutions will cause disease, and 85% of the substitutions will not cause disease. *In this case, the number of factor IX–specific residues will approach zero as progressively more factor IX sequences are added. Thus, the additional factor IX sequences will increase the fraction of causative mutations occurring at nonconserved residues.*
3. Missense mutations at 15% of the factor IX–specific residues cause hemophilia B, and the remaining 85% cause an as yet undefined disease such as a hypercoagulability that might perhaps be lethal prenatally. *In this case, the 141 factor IX–specific residues should remain conserved if other factor IX sequences are added.*

To help distinguish between these possibilities, sequence data were analyzed from the C-terminal segment (the activation and catalytic domain) of an additional five species of factor IX (Sarkar et al. 1990). An alignment of all nine species (C-terminal 9) versus the alignment of four species (C-terminal 4) indicates that 27 residues (13%) are now reclassified as nonconserved (table 4A). None of these reclassified residues were generic in the initial C-terminal 4 alignment. In contrast to the prediction of model 3, 12% of the partially generic and 21% of the factor IX–specific residues were reclassified as nonconserved (table 4B). More important, no missense mutations have been reported to cause hemophilia B in any of the 27 reclassified nonconserved residues. If there had been a corresponding decline in the number of missense mutations at conserved residues, as predicted by model 2, an additional 3.5 mutations would be expected at nonconserved residues (table 4B). However, none were observed ($P < .002$). While the data best support the predictions of model 1, more cross-species sequences and a larger sample of mutations are necessary to eliminate a hybrid model that contains some contribution from model 2 and/or model 3.

The pattern of transitions at CpG dinucleotides further supports model 1. Since the mutation rate in factor IX is much higher at CpG dinucleotides than at non-CpG dinucleotides (Koeberl et al. 1989; Green et al. 1990), the mammalian factor IX sequences at CpG

Table 4

Amino Acid Conservation and Missense Mutations in Activation and Catalytic Domains of Factor IX at Non-CpG Residues in C-terminal 4 versus C-terminal 9

A. Amino Acid Conservation in C-terminal 4 vs. C-terminal 9

	Total Conserved		Total Nonconserved	
	C-terminal 4	C-terminal 9	C-terminal 4	C-terminal 9
No. of residues	218	191	54	81
No. of missense mutations observed	48	48	1	1

B. Conserved C-terminal 4 Amino Acids Converted to Nonconserved in C-terminal 9

	Generic	Partially Generic	Factor IX Specific	Total
No. of residues converted[a]	0	10	17	27
Residues converted in each C-terminal 4 conserved class	0%	12%	21%	12%
No. of observed missense mutations expected to become C-terminal 9 nonconserved[b]	0	4.1	2.6	6.7

Note.—Conservation is defined as in Methods. Conservation of the carboxy segment of factor IX was determined on the basis of an alignment of either C-terminal 4 or C-terminal 9. Both alignments utilize sequence from human, mouse, cow, and dog. The C-terminal 9 alignment also includes sequence from sheep, pig, rat, guinea pig, and rabbit (Sarkar et al. 1990) (see Appendix A). Similar alignments have been published elsewhere (Sarkar et al. 1990; Wu et al. 1990).

[a] Residues conserved in the C-terminal 4 alignment that convert to nonconserved residues in the C-terminal 9 alignment.

[b] Number of observed missense mutations expected, on the basis of model 2, to become C-terminal 9 nonconserved; this number is calculated by multiplying the percent of C-terminal 4 conserved residues that convert to nonconserved in the C-terminal 9 alignment by the relative probability of missense mutation causing hemophilia B (table 3). For partially generic residues, 10 × .41 = 4.1, and for factor IX–specific residues 17 × .15 = 2.6.

dinucleotides should rapidly mutate at nonconserved amino acids. Therefore, the pattern observed in mammalian sequences at CpG dinucleotides should be *analogous to the pattern of non-CpG conserved dinucleotides that would be observed in more diverged species*. Transitions at the 15 conserved CpG nucleotides should cause a missense mutation resulting in disease (table 5). Transitions at 12 of the 15 possible CpG sites have been observed. These transitions have occurred in three of four generic residue sites, in five of six partially generic residue sites, and, most important, in four of five factor IX–specific residue sites. Thus, there is an almost perfect correlation between evolutionary conservation and missense mutations causing hemophilia B. Furthermore, there are four arginine residues (codons CGX, where X is any base except A) in which transitions at either C or G will produce a missense mutation. Mutations at all six evolutionarily conserved sites have been observed to cause hemophilia B, while neither of the two possible transitions have been observed in the one nonconserved CGX arginine residue (R^{403}).

Percent of Missense Mutations That Cause Hemophilia B

Factor IX, factor X, factor VII, and protein C diverged about 450–500 million years ago (Doolittle and Feng 1987). If a residue is identical in these proteases despite such a long period of evolutionary time, it is very likely to be absolutely essential for protein function. The essential nature of such generic residues (102 total) is supported by an analysis of the residues conserved in the C-terminal 4 of factor IX versus those in the C-terminal 9 of factor IX (table 4B). Generic residues are absent from the group of 27 amino acids that, as a result of the C-terminal 9 alignment, have been reclassified as nonconserved (table 4B). Additional support comes from an alignment of human and bovine factor VII, factor X, and protein C. One hundred six residues are identical in these proteins. If the four species of factor IX are added to the alignment, virtually all (96%) of these generic residues remain identical, despite an additional 450+ million years of evolutionary divergence. Thus, we conclude that most, if not all, possible missense mutations at generic residues will cause disease. From the relative frequencies of mutations in each class, it can be estimated that only 40% of all possible missense changes in factor IX will cause hemophilia B (table 3). In the context of model 1, the estimate implies that 40% of factor IX residues are important for function and that most, if not all, missense mutations in these residues will cause disease.

Discussion

Prediction of Missense Mutations That Cause Hemophilia

We have analyzed (*a*) amino acid changes that disrupt factor IX function in hemophiliacs and (*b*) amino

Table 5

Missense Changes in Factor IX That Are Due to Transitions at CpG Dinucleotides, as Function of Amino Acid Conservation

	Amino Acid Conservation Corresponding to C or G Nucleotide at CpG[a]					
	Generic	Partially Generic	Factor IX Specific	Total Conserved	Nonconserved	Total
A. No. of C or G nucleotides at CpG[b]	4	6	5	15	6	21
B. No. of our independent missense mutations[c]	2	7	5	14	0	14
C. No. of sites at which missense transitions have been reported[d]	3	5	4	12	0	12

[a] At certain arginine residues (CGX), a transition at either C or G will cause a missense mutation, while in other cases transitions at only G will cause a missense mutation. Thus, each transition which causes a missense mutation is counted separately.

[b] Nucleotides in which transitions result in missense mutations.

[c] Data include our independent transitions resulting in missense mutations at CpG dinucleotides of factor IX. Multiple mutations at the same site were judged as independent only if the haplotype differed or if a germ line of origin could be determined (table 1; Bottema et al. 1990*a*; Koeberl et al. 1990).

[d] Data are from table 1 and from Bottema et al. (1990*a*), Giannelli et al. (1990), and Koeberl et al. (1990).

acid changes that are compatible with normal factor IX function in different species. Four classes of factor IX amino acids were defined on the basis of the extent of evolutionary conservation. We document the functional importance of most, if not all, of the generically conserved residues. However, residues uniquely conserved in factor IX are sixfold less likely to give rise to hemophilia B. Thus, many mutations at factor IX–specific residues should be neutral variants. One such neutral variant in a factor IX–specific residue has recently been discovered: histidine^{257} is nonconservatively changed to tyrosine without causing hemophilia B in a male (Montandon et al. 1990).

Both (1) the relationship between missense mutation and amino acid conservation class as defined by the carboxy segment of factor IX in C-terminal 4 versus that in C-terminal 9 and (2) an analysis of missense mutations at CpG dinucleotides suggest that about 40% of the residues in factor IX are crucial for function. Most of the possible missense mutations in the remaining 60% of the residues will not cause hemophilia B; these remaining residues are likely to be "spacers," i.e., residues which maintain the position of critical amino acids but whose own side chains are not crucial for function (Doolittle and Blombäck 1964). Many of these spacer residues are classified as factor IX specific because the evolutionary time separating the mammalian factor IX sequences is insufficient to have changed many nonessential residues. The conclusions predict that nonmammalian factor IX sequences should convert many of these nonessential residues to the nonconserved class and substantially increase the likelihood that a mutation in the remaining factor IX–specific residues will cause disease.

Saturation in vitro mutagenesis of selected residues in factor IX and expression in tissue culture could help confirm that factor IX function will be significantly compromised by the amino acid substitutions found in hemophiliacs. However, the generation, confirmation, and characterization of such a large number of mutations would require many years of effort. In addition, the interpretation of data indicating that a mutation retains functional integrity is confounded by the simplicity of a cell culture system in comparison with the intact organism (i.e., mutants that score as functional in cell culture could still cause hemophilia B in humans). It would be preferable to generate a battery of transgenic mice with hemophilia B, but both the absence of a mouse model for hemophilia B and the difficulties in generating a large number of transgenic mice with independent mutations pose major technical challenges. We conclude that sequencing factor IX in more nonmammalian species and delineating the mutations in a larger sample of hemophiliacs is currently the best way to determine which missense mutations will result in disease.

The generality of the present findings can be assessed by examining the correlation between evolutionary conservation and mutations causing other severe X-linked diseases. Hemophilia A would be a good choice, because factor VIII belongs to a different gene family and more than 100 missense mutations will soon be available. Autosomal genes such as α- and β-globin are *not* good candidates for assessing the relationship between evolutionary conservation and missense mutations that disrupt function. In the cases of α- and β-globin, the marked overrepresentation of dominant mutations, heterozygote advantage, founder effect, and the biased methods of patient ascertainment pose major problems in the interpretation of the data. As an example of the problems, the aggregate mutational data in globin erroneously indicated that CpG was *not* a hot spot of mutation (Vogel and Motulsky 1986).

Mutant Analyses in Other Genes

The data, albeit meager, from saturation in vitro mutagenesis in other systems is compatible with the notion that, if one missense mutation at a residue disrupts function, then the other possible missense mutations are also very likely to disrupt function. In *Escherichia coli,* saturation mutagenesis of evolutionarily conserved residues in the region of the β-lactamase active site revealed that 14 of the 19 possible amino acid substitutions retained appreciable activity toward the penicillins (Schultz and Richards 1986). However, in all but two of the substitutions the limited characterization performed was sufficient to reveal major reductions in catalytic specificity and/or thermal stability, strongly suggesting that all these mutants would be at a selective disadvantage in vivo. Moreover, in a follow-up study using saturation mutagenesis at five codons, partial activity could commonly be found, but no mutant protein had the catalytic specificity and thermal stability of a wild-type protein (Dube and Loeb 1989). Finally, in NIH 3T3 cells, a study of substitutions at the conserved glycine^{12} of the Harvey ras protein indicated that 18 of 19 amino acid substitutions produced a transformed phenotype (Seeburg et al. 1984).

Analysis of the N-terminal segment of the lambda repressor by cassette mutagenesis indicates both some residues in which only the wild-type side chain is acceptable and other residues in which either a few or many substitutions are acceptable (Reidhaar-Olson and Sauer 1988; Lim and Sauer 1989). However, interpretation of the data is complicated by (1) the generation of multiple potentially compensatory mutations by cassette mutagenesis, (2) the biases associated with a mutation method which relies on equal pairing of inosine with all bases, (3) the use of only the N-terminal fragment of the repressor, and (4) the use of a selection or screening scheme without knowledge of how that translates into the fitness of the viral protein in its ecosystem.

Caveats

Multiple base substitutions at a single codon are rare, and no such missense mutations have yet been reported in the factor IX gene. If the present pattern is representative of the past, a residue will be conserved through evolution if the five to eight possible single-base missense changes all cause disease. It is conceivable that disease will not be caused by missense changes that arise from multiple base substitutions at a codon. However, this seems unlikely because 94% of the generic residue sites do not tolerate even highly conservative substitutions which usually involve a single-base change (see Methods and fig. 1). Therefore, the more drastic missense changes that commonly result from substitutions at two and three bases are unlikely to be tolerated.

A second caveat concerns the virtual certainty that at least a small fraction of residues may fit model 2 and perhaps model 3. Such residues will limit the extent to which evolutionary conservation can predict which mutations will cause hemophilia B in humans. Both identification of more missense mutations and additional sequencing of factor IX from nonmammalian species should ultimately allow an estimate of the fraction of residues fitting models 2 and 3.

Possible Implications for Clinical Research

The development of rapid PCR-based methods for direct sequencing and screening assures that many protein sequence variants will be detected by the analysis of DNA. Some of these variants will be found in individuals who also carry a normal allele. How does one assess the likelihood of the change being neutral, as opposed to a change that either predisposes to a multifactorial disease in heterozygotes or causes a recessively inherited disease in homozygotes? Given the expense and effort of clinical studies, it would be useful to have criteria for estimating the likelihood that a missense mutation observed in a heterozygote will produce a dysfunctional protein. If further data were to show that the present observations are generally true, the level of evolutionary conservation might provide such criteria.

Acknowledgments

We thank Joslyn Cassady, Renee McGovern, and Amy Groszbach for contributing to the sequencing; Charyl Dutton, Christine Paynton, and Gobinda Sarkar, Ph.D., for contributing to the haplotype analysis; Darrell Ricke and Paul Axelsen, M.D., for help with the conservation data alignment; and Mary Johnson for excellent secretarial assistance. We thank Karin Blakemore, M.D., and Virginia Corson, M.S., both of John Hopkins Hospital; John Rinehart, M.D., and Pat Burkey, R.N., both of Ohio State University Hospital; Joan Gill, M.D., and Luann Weik, M.S., both of Children's Hospital of Wisconsin; Roger Edson, M.D., and Bonnie LeRoy, M.S., both of University of Minnesota Hospital; Charles Sexauer, M.D., and Joyce Shissler, R.N., both of Oklahoma Children's Memorial Hospital; Marian Koerper, M.D., and Pat Liddy, R.N., both of University of California, San Francisco Hemophilia Center; Carol Kaspar, M.D., and Peter Mockary, M.S., both of Orthopedic Hospital Hemophilia Center of Los Angeles; Bruce Furie, M.D., and Steven Limentani, M.D., both of New England Medical Center; and Emily Czapek, M.D., and DeAnne Schmid, R.N., both of Regional Comprehensive Hemophilia Center of Central and Northern Illinois, for providing pedigrees and blood samples. We thank Kim Maness and Melinda Cohen of Vanderbilt University for providing DNA samples, laboratory values, and pedigrees. This work was aided by grant 5-647 from the March of Dimes—Birth Defects Foundation and by NIH grant HL39762. We thank Paul Axelsen, M.D., Whyte Owen, Ph.D., and David Fass, Ph.D., for helpful discussions.

Appendix A

Alignment of N-Terminal Segment of Factor IX

```
                                     *Intron 1
                     -30v      -20v      -10v       +1v
Human          MAESPGLITICLLGYLLSAECTVFLDHENANKILNRPKRYNSG
Cow                                                   ....
Mouse          ....RA....F.......T..A....R...T...T........
Dog            ...AS................A....R...T...S........
Rat            ..DAP..
Macaque        .......
```

```
                                                *Intron 2*Intron 3
                  10v       20v       30v       40v       50v
Human          KLEEFVQGNLERECMEEKCSFEEAREVFENTERTTEFWKQYVDGDQCESN
Cow            ......R.......K.................K.................
Mouse          ......R.......I..R..............K.................
Dog            ......R.......I..R..............K.................
```

```
                                             *Intron 4
                  60v       70v       80v       90v      100v
Human          PCLNGGSCKDDINSYECWCPFGFEGKNCELDVTCNIKNGRCEQFCKNSAD
Cow            ......M............QA....T.....A..S......K....RDT.
Mouse          ......I.....S......QV....R.....A.........K......P.
Dog            ....D.V............RA....................K....LGP.
```

```
                                      *Intron 5
                 110v      120v      130v
Human          NKVVCSCTEGYRLAENQKSCEPAVPFPCG
Cow            ........D......D.............
Mouse          ...I.......Q...D......T......
Dog            ........T..Q...D.R...........
```

Figure A1 Amino acid sequences of factor IX from mammalian species, aligned from published sequences. The full-length proteins were aligned from human (Yoshitake et al. 1985), cow (Katayama et al. 1979), dog (Evans et al. 1989), and mouse (Wu et al. 1990). For the activation and catalytic domains, additional sequence from sheep, guinea pig, rat, pig, and rabbit (Sarkar et al. 1990) was used for the analyses that compared nine with four species (table 4). The translation start site is based on additional data from rat and macaque (Pang et al. 1990). The factor IX–specific residues are underlined.

```
                     140v      * 150v      160v                170v
Human             RVSVSQTSK-LTR AEAVFPDVDYVNSTEA----------ETILDNITQST
Sheep             .A..LH...K... ..TI.SNMN.E..S..----------.I.W..V...N
Pig                  ..HSPTT... ..II.SNM..E....V----------.P...SL.E.N
Rabbit            .....HA..KI.. .TTI.SNTE.E.F...----------...RG.V..RS
Guinea Pig        ...IPSV..EHN. .N.I.SRMG...F.DDETIWDDNDDD...W..S.E..
Rat               ....AYN..KI.. ..T..SNT..G....L--ILDDITN-S.....L.ENS
Mouse             .A.I.YS..KI.. ..T..SNM..E.....VFIQDDITD-GA..N.V.E.S
Cow               .....HI..K... ..TI.SNTN.E..S..----------.I.W..V...N
Dog               ....PHI.MTR.. ..TL.SNM..E....V----------.K....V..--

                                      *Intron 6
                      180v*      190v      200v      210v      220v
Human             QSFNDFTR VVGGEDAKPGQFPWQVVLNGKVDAFCGGSIVNEKWIVTAAHC
Sheep             ...D..N. .......AR.......L.H.EIA............V......
Pig               ..SD..I. I....N..........L....I........I....V......
Rabbit            ..SD.... I....N..........L.....E.......I....V......
Guinea Pig        KPSDE.F. ................L...ETE...................
Rat               EPI..... .....N.....I....I...EIE.....A.I...........
Mouse             E.L..... .....N.....I....I...EIE.....A.I...........
Cow               ...DE.S. .......ER.......L.H.EIA............V......
Dog               -PL..... ....K...........L.............I....V......

                          *Intron 7
                      230v      240v      250v      260v      270v
Human             VETGVKITVVAGEHNIEETEHTEQKRNVIRIIPHHNYNAAINKYNHDIAL
Sheep             IKP............T.KP.P.........A..Y.G...S....S.....
Pig               I.P..........Y.T....P...R.....A....S...TV...S.....
Rabbit            IKPDDN.......Y..Q...N............Y.K...T..........
Guinea Pig        ILP.I..E....K....KK.D...R...TQ..L..S...SF...S.....
Rat               LKP.D..E........DEK.D...R.....T....Q...T....S.....
Mouse             LKP.D..E.....Y..DKK.D...R.....T....Q...T....S.....
Cow               IKP............T.KP.P.........A..Y.S...S....S.....
Dog               I.PD....I......T.KR...........T.L..S...T..........

                      280v      290v      300v      310v      320v
Human             LELDEPLVLNSYVTPICIADKEYTNIFLKFGSGYVSGWGRVFHKGRSALV
Sheep             .......E............R..........Y..........NR....SI
Pig               .......T..................................NR....TI
Rabbit            ....K..T...........NR.......N.............NR..Q.SI
Guinea Pig        ....K..S...........NR..........A.......KL.SQ..T.SI
Rat               ....K..I.........V.N...................K..N...Q.SI
Mouse             ....K..I.........V.NR..................K..N...Q.SI
Cow               .......E............RD.....S...Y.......K..NR....SI
Dog               .......T............R..S..................N.....SI
```

Appendix A (continued)

```
                    330v      340v      350v      360v      370v
Human           LQYLRVPLVDRATCLRSTKFTIYNNMFCAGFHEGGRDSCQGDSGGPHVTE
Sheep           ....K...................H.....Y....K..............
Pig             ....K..............V...S...........K...L..........
Rabbit          .......F.......................DV..K...E..........
Guinea Pig      ..................................................
Rat             ....................S.........YR...K...E..........
Mouse           ..................T...........YR...K...E..........
Cow             ....K...............S..SH.....Y....K..............
Dog             ....K..............................K..............


                    380v      390v      400v       410v
Human           VEGTSFLTGIISWGEECAMKGKYGIYTKVSRYVNWIKEKTKLT
Sheep           ................................
Pig             ..................V................
Rabbit          ..................I.....V..R..W....
Guinea Pig      ....N..............................
Rat             ...................................
Mouse           ...........................................
Cow             ...........................................
Dog             ...I.......................................
```

Appendix B

Alignment of Related Coagulation Serine Proteases

```
                    -39                        * Intron 1
Human Factor IX        MAE SPGLITICLL GYLLSAECTV FLDHENANKI LNRPKRYNSG
Bovine Factor IX                                                  ....
Human Factor X       MGRPL HLV.LSAS.A .L..LG.-SL .IRR.Q..N. .A.VT.A..-
Bovine Factor X      MAGLL HLV.LSTA.G .L.RPAG-S. ..PRDQ.HRV .Q.AR.A..-
Human Factor VII        MV .QA.RLL... LG.QGCLAA. .VTQ.E.HGV .H.RR.A.A-
Bovine Factor VII                                                 A.-.
Human Protein C     MWQLTS LLLFVATWGI SGTPAPLDS. .SSS.R.HQV .RIR..A..-
Bovine Protein C        TS LLLFVT.WGI SSTPAPPDS. .SSS.R.HQV .RIR..A..-


                5                                   *Intron 2*Intron 3
H. Factor IX   KLEEFVQGNL ERECMEEKCS FEEAREVFEN TERT TEFWKQ YVDGDQC----ESN
B. Factor IX   ......R... ....K..... .......... ..K. ...... .......----...
H. Factor X    F...MKK.H. .......T.. Y........D SDK. N...NK .K.....----.TS
B. Factor X    F...VK.... ....L..A.. L........D A.Q. D...SK .K.....----.GH
H. Factor VII  F...LRP.S. ....K..Q.. ......I.KD A... KL..IS .S.....----A.S
B. Factor VII  F...LRP.S. ....R..L.. ....H.I.R. E... RQ..VS .N.....----A.S
H. Protein C   F...LRHSS. ....I..I.D ....K.I.Q. VDD. LA..SK H......LVLPLEH
B. Protein C   F...LRP..V ....S..V.E ......I.Q. ..D. MA..SK .S.....EDRPSGS


             55                                     * Intron 4
H. Factor IX   PCL----NGGSCKD DINSYECWCP FGFEGKNCEL DVT----CNIKNGR CEQFCKNSAD
B. Factor IX   ...----...M... .........Q A....T.... .A.----.S..... .K....RDT.
H. Factor X    ..Q----.Q.K... GLGE.T.T.L E......... FTRK--L.SLD..D .D...-HEEQ
B. Factor X    ...----.Q.H... G.GD.T.T.A E........F STRE--I.SLD..G .D...-REER
H. Factor VII  ..Q----....... QLQ..I.F.L PA...R...T HKDDQLI.VNE..G ...Y.SDHTG
B. Factor VII  ..Q----.....E. QLR..I.F.. D....R...T .KQSQLI.AND..G ...Y.GADPG
H. Protein C   ..ASLCCGH.T.I. G.G.FS.D.R S.W..RF.QR E.SF-LN.SLD..G .THY.-LEEV
B. Protein C   ..DLPCCGR.K.I. GLGGFR.D.A E.W..RF.LH E.RF-SN.SAE..G .AHY.-MEEE


            105                         * Intron 5
H. Factor IX   NKVVCSCTEG YRLAENQKSC EPAVPFPCGR VSVSQTSK-LT RAEAVFPDVD YVNSTEAETI
B. Factor IX   ........D. .....D.... .......... ....HI..K.. ...TI.SNTN .E..S...I.
H. Factor X    .S.....AR. .T..D.G.A. I.TG.Y...K QTLERRKRSVA Q.TSSSGEAP DSITWKPYDA
B. Factor X    SE.R...AH. .V.GDDS.S. VSTER....K FTQGR---SSR W.IHTSEDAL DASEL.HYDP
H. Factor VII  T.RS.R.H.. .S.LADGV.. T.T.EY...K IPILEKR---- ---------- ----------
B. Factor VII  AGRF.W.H.. .A.QADGV.. A.T.EY...K IP.LEKR---- ---------- ----------
H. Protein C   GWRR...AP. .K.GDDLLQ. H...K..... PWKRMEK.RSH LKRDTE---- ----------
B. Protein C   GRRH...AP. ...EDDHQL. VSK.T..... LGKRMEK.RK. LKRDTN---- ----------
```

Figure B1 Factor IX sequences aligned with human and bovine sequences from related coagulation serine proteases factor VII (Hagen et al. 1986; Takeya et al. 1988), factor X (Fung et al. 1984, 1985), and protein C (Long et al. 1984; Beckmann et al. 1985). The generic and partially generic residues are underlined.

Appendix B (continued)

```
            165                                                    *Intron 6
H. Factor IX    LD------------NITQSTQS FN--DFTRVVGG EDAKPGQFPW Q VVL-NGKVDA
B. Factor IX    W.-------------.V...N.. .D--E.S..... ...ER..... . .L.-H.EIA.
H. Factor X     DLDPTENPFDLLDF.Q..PERG D.--NL..I... QEC.D.EC.. . AL.I.EENEG
B. Factor X     DLSPTESSLDLLGL.R.EPSAG EDGSQVV.I... R.CAE.EC.. . AL.V.EENEG
H. Factor VII   --------------------NA SK--PQG.I... KVCPK.EC.. . .L.-LVNGAQ
B. Factor VII   --------------------NG SK--PQG.I... HVCPK.EC.. . AM.-KLNGAL
H. Protein C    -------------------D.E DQ--VVP.LID. KMTRR.DS.. . ...LDS.KKL
B. Protein C    -----------------.VD.K DQ--LVP.I.D. QE.GW.ES.. . A..LDS.KKL


            205                                *Intron 7
H. Factor IX    FCGGSIVNEK WIVTAAHC---VE TGVKITVVAG EHNIEETEHT EQKRNVIRII
B. Factor IX    .......... .V......---IK P......... ...T.KP.P. ........A.
H. Factor X     ....T.LS.F Y.L.....---LY QAKRFK.RV. DR.T.QE.GG .AVHE.EVV.
B. Factor X     ....T.L..F YVL.....---LH QAKRF..RV. DR.T.QE.GN .MAHE.EMTV
H. Factor VII   L...TLI.TI .V.S....FDKIK NWRNLIA.L. ..DLS.HDGD ..S.R.AQV.
B. Factor VII   L...TL.GPA .V.S....FERLR SRGNL.A.L. ..DLSRV.GP ..E.R.AQ..
H. Protein C    A..AVLIHPS .VL.....---MD ESK.LL.RL. .YDLRRW.KW .LDLDIKEVF
B. Protein C    V..AVLIHVS .VL.V...---LD SRK.LI.RL. .YDMRRW.SW .VDLDIKEV.


            255
H. Factor IX    PHHNYNAAIN KYNHDIALLE LDEPLVLNSY VTPICIADKE YTNIFL--KFGS
B. Factor IX    .Y.S...S.. ..S....... .....E.... ........RD .....S--...Y
H. Factor X     K.NRF--TKE T.DF...V.R .KT.ITFRMN .A.A.LPERD WAEST.-MTQKT
B. Factor X     K.SRF--VKE T.DF...V.R .KT.IRFRRN .A.A.LPE.D WAEAT.-MTQKT
H. Factor VII   IPST.--VPG TT.......R .HQ.V..TDH .V.L.LPERT FSERT.-AFVRF
B. Factor VII   VPKQ.--VPG QTD..V...Q .AQ.VA.GDH .A.L.LP.PD FADQT.-AFVRF
H. Protein C    V.P..--SKS TTDN.....H .AQ.AT.SQT IV...LP.SG LAERE.NQAGQE
B. Protein C    I.P..--TKS TSDN.....R .AL.AT.SQT IV...LP.SG LSERK.TQVGQE


            305
H. Factor IX    GYVSGWG----RVF HKGRS-ALVLQ YLRVPLVDRA TCL-----RSTKFTI YNNMFCAGFH
B. Factor IX    .......----K.. NR...-.SI.. ..K....... ...-----.....S. .SH.....Y.
H. Factor X     .I...F.----.TH E...Q-STR.K M.E..Y...N S.K-----L.SS.I. TQ......YD
B. Factor X     .I...F.----.TH E...L-SST.K M.E..Y...S ..K-----L.SS... TP......YD
H. Factor VII   SL.......----QLL DR.AT-..E.M V.N..RLMTQ D..QQSRKVGDSPN. TEY.....YS
B. Factor VII   SA.......----QLL ER.VT-.RK.M VVL..RLLTQ D..QQSRQ.PGGPVV TD......YS
H. Protein C    TL.T...YHSS.EK EAK.NRTF..N FIKI.V.PHN E.S-----EVMSNMV SE..L...IL
B. Protein C    TV.T...----YRD ETK.NRTF..S FIK..V.PYN A.V-----HAMENK. SE..L...IL


           355
H. Factor IX    EGGRDSCQGD SGGPHVTEVE GTSFLTGIIS WGEECAMKGK YGIYTKVSRY VNWIKEKTKL T
B. Factor IX    ...K...... .......... .......... .......... .......... .......... .
H. Factor X     TKQE.A.... .......RFK D.Y.V...V. ...G..R... .......TAF LK..DRSM.T R
B. Factor X     TQPE.A.... .......RFK D.Y.V...V. ...G..R... F.V.....NF LK..DLIM.A R
H. Factor VII   D.SK...K.. .....A.HYR ..WY....V. ..QG..TV.H F.V..R..Q. IE.LQKLMRS E
B. Factor VII   D.SK...K.. .....A.RFR ..W....VV. ...G..AA.H F....R.... TA.LRQLMGH P
H. Protein C    GDRQ.A.E.. ....M.ASFH ..W..V.LV. ...G.GLLHN ..V....... LD..HGHIRD K
B. Protein C    GDP..A.E.. ....M..FFR ..W..V.LV. ...G.GRLYN ..V....... LD..YGHI.A Q
```

References

Beckmann RJ, Schmidt RJ, Santerre RF, Plutzy J, Crabtree GR, Long GL (1985) The structure and evolution of a 461 amino acid human protein C precursor and its messenger RNA, based upon the DNA sequence of cloned human liver cDNAs. Nucleic Acids Res 13:5233–5247

Bottema CDK, Bottema MJ, Ketterling RP, Yoon H-S, Janco RL, Phillips III JA, Sommer SS (1991) Why does the human factor IX gene have a G + C content of 40%? Am J Hum Genet 49:839–850

Bottema CDK, Ketterling RP, Yoon H-S, Sommer, SS (1990*a*) The pattern of factor IX germ-line mutation in Asians is similar to that of Caucasians. Am J Hum Genet 47:835–841

Bottema CDK, Koeberl DD, Ketterling RP, Bowie EJW, Taylor SAM, Lillicrap D, Shapiro A, et al (1990*b*) A past mutation at isoleucine^{397} is now a common cause of moderate/mild haemophilia B. Br J Haematol 75:212–216

Bottema CDK, Koeberl DD, Sommer SS (1989) Direct carrier testing in 14 families with hemophilia B. Lancet 2: 526–529

Camerino G, Grzeschik KH, Jaye M, DeLaSalle H, Tolstoshev P, Lecocq JP, Heilig R, et al (1984) Regional localization on the human X chromosome and polymorphism of the coagulation factor IX gene (hemophilia B locus). Proc Natl Acad Sci USA 81:498–502

Doolittle RF, Blombäck B (1964) Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. Nature 202:147–152

Doolittle RF, Feng DF (1987) Reconstructing the evolution of vertebrate blood coagulation from a consideration of the amino acid sequences of clotting proteins. Cold Spring Harbor Symp Quant Biol 52:869–874

Dube DK, Loeb LA (1989) Mutants generated by the insertion of random oligonucleotides into the active site of the β-lactamase gene. Biochemistry 28:5703–5707

Evans JP, Watzke HH, Ware JL, Stafford DW, High KA (1989) Molecular cloning of a cDNA encoding canine factor IX. Blood 74:207–212

Eyster ME, Lewis JH, Shapiro SS, Gill F, Kajani M, Prager D, Djerassi I, et al (1980) The Pennsylvania hemophilia program 1973–1978. Am J Hematol 9:277–286

Fung MR, Campbell RM, MacGillivray RTA (1984) Blood coagulation factor X mRNA encodes a single polypeptide chain containing a prepro leader sequence. Nucleic Acids Res 12:4481–4492

Fung MR, Hay CW, MacGillivray RTA (1985) Characterization of an almost full-length cDNA coding for human blood coagulation factor X. Proc Natl Acad Sci USA 82: 3591–3595

Furie B, Furie BC (1988) The molecular basis of blood coagulation. Cell 53:505–518

Giannelli F, Green PM, High KA, Lozier JN, Lillicrap DP, Ludwig M, Olek K, et al (1990) Haemophilia B: database of point mutations and short additions and deletions. Nucleic Acids Res 18:4053–4059

Green PM, Montandon AJ, Bentley DR, Ljung R, Nilsson IM, Giannelli F (1990) The incidence and distribution of CpG→TpG transitions in the coagulation factor IX gene: a fresh look at CpG mutational hotspots. Nucleic Acids Res 18:3227–3231

Gustafson S, Proper JA, Bowie EJW, Sommer SS (1987) Parameters affecting the yield of DNA from human blood. Anal Biochem 165:294–299

Hagen FS, Gray CL, O'Hara P, Grant FJ, Saari GC, Woodbury RG, Hart CE, et al (1986) Characterization of a cDNA coding for human factor VII. Proc Natl Acad Sci USA 83:2412–2416

Katayama K, Ericsson LH, Enfield DL, Walsh KA, Neurath H, Davie EW, Titani K (1979) Comparison of amino acid sequence of bovine coagulation factor IX (Christmas factor) with that of other vitamin K–dependent plasma proteins. Proc Natl Acad Sci USA 76:4990–4994

Ketterling RP, Bottema CDK, Koeberl DD, Ii S, Sommer SS. T^{296}→M, a common mutation causing mild hemophilia B in the Amish and others: founder effect, variability in factor IX activity assays and rapid carrier detection. Hum Genet (in press)

Ketterling RP, Bottema CDK, Phillips JP III, Sommer SS (1991) Evidence that descendants of three founders comprise about 25% of hemophilia B in the United States. Genomics 10:1093–1096

Koeberl DD, Bottema CDK, Buerstedde J-M, Sommer SS (1989) Functionally important regions of the factor IX gene have a low rate of polymorphism and a high rate of mutation in the dinucleotide CpG. Am J Hum Genet 45: 448–457

Koeberl DD, Bottema CDK, Ketterling RP, Bridge PJ, Lillicrap DP, Sommer SS (1990) Mutations causing hemophilia B: direct estimate of the underlying rates of spontaneous germ-line transitions, transversions, and deletions in a human gene. Am J Hum Genet 47:202–217

Lehninger AL (1975) Biochemistry: the molecular basis of cell structure and function, 2d ed. Worth, New York

Lim WA, Sauer RT (1989) Alternative packing arrangements in the hydrophobic core of lambda repressor. Nature 339:31–36

Long GL, Belagaje RM, MacGillivray RTA (1984) Cloning and sequencing of liver cDNA coding for bovine protein C. Proc Natl Acad Sci USA 81:5653–5656

Montandon AJ, Green PM, Bentley DR, Ljung R, Nilsson IM, Giannelli F (1990) Two factor IX mutations in the family of an isolated haemophilia B patient: direct carrier diagnosis by amplification mismatch detection (AMD). Hum Genet 85:200–204

Pang C-P, Crossley M, Kent G, Brownlee GG (1990) Comparative sequence analysis of mammalian factor IX promoters. Nucleic Acids Res 18:6731–6732

Rees DJG, Jones IM, Handford PA, Walter SJ, Esnouf MP, Smith KJ, Brownlee GG (1988) The role of β-hydoxyaspartate and adjacent carboxylate residues in the first EGF domain of human factor IX. EMBO J 7:2053–2061

Reidhaar-Olson JF, Sauer RT (1988) Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science 241:53–57

Sarkar G, Koeberl DD, Sommer SS (1990) Direct sequencing of the activation peptide and the catalytic domain of the factor IX gene in six species. Genomics 6:133–143

Schultz SC, Richards JH (1986) Site-saturation studies of β-lactamase: production and characterization of mutant β-lactamases with all possible amino acid substitutions at residue 71. Proc Natl Acad Sci USA 83:1588–1592

Seeburg PH, Colby WW, Capon DJ, Goeddel DV, Levinson AD (1984) Biological properties of human c-Ha-*ras*1 genes mutated at codon 12. Nature 312:71–75

Sommer SS, Sarkar G, Koeberl DD, Bottema CDK, Buerstedde J-M, Schowalter DB, Cassady JD (1990) Direct sequencing with the aid of phage promoters. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) PCR protocols: a guide to methods and applications. Academic Press, New York, pp 197–205

Stoflet ES, Koeberl DD, Sarkar G, Sommer SS (1988) Genomic amplification with transcript sequencing. Science 239:414–419

Takeya H, Kawabata S-I, Nakagawa K, Yamamichi Y, Miyata T, Iwanaga S, Takao T, et al (1988) Bovine factor VII its purification and complete amino acid sequence. J Biol Chem 263:14868–14877

Thompson AR, Bajaj SP, Chen SH, MacGillivray RTA (1990) "Founder" effect in different families with haemophilia B mutation. Lancet 1:418

Vogel F, Motulsky AG (1986) Problems and approaches. In: Human genetics. Springer, Berlin, pp 368–371

Winship PR, Anson DS, Rizza CR, Brownlee GG (1984) Carrier detection in haemophilia B using two further intragenic restriction fragment length polymorphisms. Nucleic Acids Res 12:8861–8872

Winship PR, Rees DJG, Alkan M (1989) Detection of polymorphisms at cytosine phosphoguanadine dinucleotides and diagnosis of haemophilia B carriers. Lancet 1:631–634

Wu S-M, Stafford DW, Ware U (1990) Deduced amino acid sequence of mouse blood-coagulation factor IX. Gene 86: 275–278

Yoshitake S, Schach BG, Foster DC, Davie EW, Kurachi K (1985) Nucleotide sequence of the gene for human factor IX (antihaemophilic factor B). Biochemistry 24:3736–3750